Drug Saf (2014) 37:945–959
DOI 10.1007/s40264-014-0214-3

ORIGINAL RESEARCH ARTICLE

Fidelity Assessment of a Clinical Practice Research Datalink Conversion to the OMOP Common Data Model

Amy Matcho · Patrick Ryan · Daniel Fife · Christian Reich

Published online: 4 September 2014

Abstract

Background The unique structure and coding of the Clinical Practice Research Datalink (CPRD) presents challenges for epidemiologic analysis and for comparisons with other databases. To address this limitation we sought to transform CPRD into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

Methods An extraction, transformation and loading process was developed, which detailed source code mappings, Read code domain classification, an imputation algorithm for drug duration and special handling of lifestyle/clinical data. Completeness and accuracy of the above elements were assessed. A final validation exercise involved replication of a published case–control study that examined use of nonsteroidal anti-inflammatory drugs (NSAIDs) and the risk of first-time acute myocardial infarction (AMI) in raw CPRD data and the CPRD CDM.

Findings All elements of the CPRD CDM transformation were assessed to be of high quality. 99.9 % of database condition records and 89.7 % of database drug records were mapped (majority unmapped drugs were devices and over-the-counter products); 3.1 % of duration imputations were deemed possibly erroneous and prevalences for selected conditions and drugs across CPRD raw and CDM data were equivalent. Results between the replication raw data and CDM study agreed for conditions, demographics and lifestyle data with slight NSAID exposure data loss owing to unmapped drugs.

Conclusion CPRD can be accurately transformed into the OMOP CDM with acceptable information loss across drugs, conditions and observations. We determined that for a particular use, case CDM structure was adequate and mappings could be improved but did not substantially change the results of our analysis.

Key Points

A transformation of the Clinical Practice Research Datalink (CPRD) into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) was performed

Quality assessments indicated that source code mappings, Read code domain classification, imputation algorithm for drug duration and special handling of lifestyle/clinical data were accurate and acceptable data loss occurred across CDM domains

A case–control replication study was performed on the CPRD raw data and the CPRD CDM and results between the raw data and CDM study agreed for conditions, demographics and lifestyle data. There was slight nonsteroidal anti-inflammatory drug exposure data loss caused by unmapped drugs

Electronic supplementary material The online version of this article (doi:10.1007/s40264-014-0214-3) contains supplementary material, which is available to authorized users.

A. Matcho (✉) · P. Ryan · D. Fife
Janssen Research and Development, LLC, 920 US Route 202, Raritan, NJ 08869, USA
e-mail: AMatcho@its.jnj.com

C. Reich
AstraZeneca PLC, 35 Gatehouse Dr., Waltham, MA 02451, USA

1 Introduction

The Clinical Practice Research Datalink (CPRD), formerly known as the General Practice Research Database or GPRD, is

a population-based electronic health record (EHR) from general practices in the UK. Though it is one of the primary databases used in epidemiologic research [1–3], the unique structure and coding of the CPRD data presents challenges for analysis and for comparisons with other databases. For instance, it is difficult to construct complete code sets in CPRD because of varying terminologies for the same medical concept in their coding schema and use of lifestyle and clinical data such as laboratory tests requires manipulation of multiple tables and nested lookup files. To address these limitations and others, we sought to transform the CPRD data into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 4, which includes a standard representation of healthcare experiences, common vocabularies for coding clinical concepts, and thus facilitates comparable analysis across disparate databases [4, 5].

Efforts to transform US claims databases into the CDM have generally been successful. For example, Overhage et al. [6] transformed data from five different observational databases (a mix of US claims databases and EHR data) into separate CDM instances and concluded that they had achieved an acceptable representation of the data by examining the proportion of terms and database records for drugs and conditions that could be mapped using the common vocabularies. The percentage of database records mapped had a range of 93.2–99.7 % for conditions and 88.8–97.6 % for medications [6]. In contrast, in a recent attempt to convert The Health Improvement Network data (THIN) (a database similar in structure and content to the CPRD) to the OMOP CDM, the authors concluded that the proportion of condition and drug codes mapped was insufficient (94 % of database condition records and 75 % of condition terms mapped and 93 % of database drug records and 45 % of drug exposure terms mapped) for quality epidemiological analyses and that the THIN data structure was an impediment to a successful conversion [7].

In the present study, we performed a CPRD to CDM conversion, evaluated the accuracy of this conversion and further assessed the adequacy of the conversion by attempting to replicate a prior published study by Schlienger et al. [8] in the raw CPRD data and the CPRD CDM. The study replicated was originally performed by the Boston Collaborative Drug Surveillance Program (BCDSP), a research organization that has participated in the evaluation and quality control of the CPRD from its inception and has published a large number of papers in the area of drug safety with CPRD data [9]. In the published study, the authors assessed the relationship between exposure to nonsteroidal anti-inflammatory drugs (NSAIDS) and incident acute myocardial infarction (AMI).

2 Methods

2.1 CPRD Transformation to the OMOP CDM

For the transformation, we used the CPRD version that contained data collected through July 29, 2013 and began by designing an extraction, transformation and loading process [10].

Table 1 provides the CDM table names, descriptions and CPRD source data tables for all CDM tables [5] that

Table 1 Populated OMOP CDM tables, descriptions and CPRD source data tables in the CPRD CDM conversion

CDM table name	Description	CPRD raw data tables
Person	Demographic information about a person	Patient
Drug exposure	Association between a person and a drug at a specific time	Therapy, immunisation, clinical, referral, test
Drug era	Association between a person and a drug over a specific time period	Therapy, immunisation, clinical, referral, test
Condition occurrence	A diagnosis or condition that has been recorded about a person at a certain time	Clinical, referral, test
Condition era	A diagnosis or condition over a period of time	Clinical, referral, test
Observation period	Time intervals during which healthcare information, such as drugs, conditions, and other clinical observations, may be available	Patient, practice
Observation	Observations are clinical facts, such as laboratory tests, signs/symptoms, which are not captured within other CDM tables	Clinical, referral, test, additional
Procedure occurrence	Procedures carried out on the person	Clinical, referral, test
Visit occurrence	Visits for healthcare services of the person	Consultation
Death	Time and cause of death of the person	Patient
Provider	Information about healthcare providers	Staff
Care site	Information about the site of care	Practice

CDM common data model, *CPRD* clinical practice research datalink, *OMOP* observational medical outcomes partnership

had the equivalent data in CPRD. We sought to populate all these CDM fields with the appropriate CPRD data. Not all patients from the CPRD raw data were included within the CPRD CDM; those that met the CPRD provided definition of a valid patient for research purposes were included (met acceptability criterion and had observation time in the database). Out of 15,000,986 patients in the raw CPRD data, 11,342,669 met the definition for inclusion in the CPRD CDM or 75.6 %. Additionally, data not within the patient's valid observation period by convention are not converted over to the CDM; 23 % of drug exposures, 35 % of conditions, 27 % of procedures and 16.7 % of observations were not within the patient's valid observation period and were not included in the CPRD CDM. The overwhelming majority of CPRD data not within valid observation time are medical history data, data that are prior to the patient joining the practice or prior to the practice data being classified as 'up-to-standard'. One notable omission of CPRD source data from the CPRD CDM is referral information such as specialty and urgency (referral conditions, procedures and observations and their event dates were captured). In the following, we report mapping difficulties, imputation required, or structural differences between the two data sources we encountered.

2.1.1 Multilex to RxNorm Mappings

RxNorm, a US-based normalized naming system for generic and branded drugs, is the standard drug lexicon used in the CDM. CPRD uses Multilex codes to identify medication. All content (e.g. conditions, procedures) in the OMOP CDM are referred to by concepts. The OMOP Standard Vocabularies are used to understand and make use of these concepts. We assigned an RxNorm concept to each Multilex source code using mappings from version 4.3 of the OMOP Standard Vocabularies for the CPRD CDM conversion [11]. These mappings used the Multilex code components that identify ingredient, strength and form of each drug and then constructed a mapping to equivalent RxNorm components. A full mapping to an RxNorm product was established if all components could be mapped and a product in RxNorm existed with the same combination of ingredient, strength and form. In the case where a product available in the UK was not available in the US (because of strength or formulation differences) or an ingredient approved by the European Medicines Agency (EMEA) has not been approved by the US Food and Drug Administration (FDA), a new Multilex concept with all the attributes of a RxNorm concept (ingredient, strength and form) was created for inclusion in the OMOP standard dictionaries.

Validation efforts for drug exposure mappings included generating the proportion of database records in the CPRD CDM drug exposure file mapped to RxNorm concepts and proportion of terms mapped (Multilex codes that exist in the raw CPRD data) to RxNorm concepts. In addition, the top 100 most frequently occurring therapies in CPRD were reviewed for mapping completeness and accuracy and the top 100 unmapped therapies were evaluated to determine if mappings were in fact possible for these high-frequency codes. To test the theory that other avenues of CDM information loss may occur besides mapping losses, drug prevalences for all database years between the CPRD raw data and the CPRD CDM were compared. We wrote a SQL program to estimate prevalence against the raw schema, and independently wrote a different SQL program to estimate prevalence against the CDM. Because Multilex to RxNorm mappings are 1:1 in only 13 % of cases, we applied the Multilex to RxNorm mappings to the CPRD raw data to Multilex codes that occurred during valid patient observation time. We only included patients in the raw data prevalences that were acceptable patients and had valid observation time and drug exposure dates needed to be within the patient's observation period.

2.1.2 Ingredient Information for Drug Products

In the OMOP Standard Vocabularies, RxNorm clinical drug (drug product) concepts that contain strength and formulation information have relationships with ingredient concepts. This allows drug exposure data to be aggregated based on ingredients to create drug eras, which can be described as an inferred period of continuous exposure to a certain ingredient over a certain period of time with a 30-day persistence window (duration allowed between subsequent drug records) [12]. It is important to note that the OMOP CDM applies a standard convention for deriving drug eras based on a 30-day persistence window and this convention is applied consistently across all databases. However, if a specific analysis use case requires a different set of assumptions for inferring consistent episodes of exposure, the CDM can accommodate this with the drug exposure table. If a clinical drug to ingredient relationship is not provided in RxNorm, then that drug was not included in the drug era file, which contains drug eras as described above. To assess the impact of this on the CPRD to CDM conversion, we evaluated the percentage of CPRD CDM drug exposure clinical drug records that had no ingredient relationship, and the proportion of drug exposures affected.

2.1.3 Drug Exposure Duration Imputation

In the CPRD data, prescription duration is not a required field, and only 7 % of drug exposures are recorded with a duration value. Drug quantity is recorded more consistently (99.3 % of all drug exposures have a valid quantity value)

and a CPRD-derived numeric daily dose field is provided for drug exposures. However, 26 % of numeric daily dose values are invalid, primarily for prescriptions with instructions to take 'as needed' or 'as directed' with no clear daily dose indicated. In addition, medications such as inhalers may have an amount of containers given in the quantity field that will not yield a valid duration when divided by numeric daily dose. Thus, we imputed drug duration for all exposures with invalid duration values of 0 (93 % of data) or >365 (0.0004 % of data) days. We performed the imputation stepwise:

1. If the CPRD duration field was invalid, we used the most common valid duration in the data for that combination of product, numeric daily dose, quantity, and number of packs given.
2. If such a combination did not produce a valid duration value in the data, then the most common valid duration in the data for the product only was used.
3. Last, if there were no valid durations in the data for a particular product, we set the duration to 1 day.

For validation purposes, we identified and examined problematic imputations after filtering out credible records with durations of 28 or 30, those with absolute difference no greater than 5 between quantity/numeric daily dose and imputed duration, and numeric daily dose = 0 (implies duration will be difficult to assess). We also examined separately database records with a valid numeric daily dose (>0) to calculate proportions of database records with imputed duration equivalent to quantity/numeric daily dose and proportions of database records with absolute difference no greater than 5 between quantity/numeric daily dose and imputed duration.

2.1.4 CDM Domain Classification Efforts

The Read dictionary version 2 is a coded thesaurus of clinical terms, in use in the UK National Health Service (NHS) to capture all aspects of patient care, including diagnoses, symptoms, findings, procedures, laboratory tests and care administration. This contrasts with coding systems in US claims databases that typically provide separate dictionaries for diagnoses and procedures and/or a way to distinguish between the two. In addition, US claims databases generally place codes for procedures and diagnoses in separate fields while Read codes are placed in one field with no domain information provided from the data structure or the Read code itself. Therefore, a domain classification effort for all Read codes was necessary to partition Read code records into the appropriate condition, procedure and observation CDM domains.

A method making use of the hierarchical nature of the Read dictionary was devised to perform this partition. Read codes are comprised of five hierarchical levels, with a higher level functioning as the 'parent' of the next lower 'child' level. The first level contains the Read chapter that provides a crude indication of domain (e.g. Read chapters A–Z usually indicate conditions, 7 indicates procedures). Though there were multiple domain types within chapters, the first four levels could be used to identify domains systematically. Therefore, all Read codes with the same values in the first four levels were reviewed manually by a clinician and classified to the same CDM domain in the OMOP Standard Vocabularies. To validate this method, the 100 most frequently occurring conditions, procedures and observations in CPRD were reviewed for domain classification accuracy.

2.1.5 Read to SNOMED-CT Mappings

In the CDM, the systematized nomenclature of medicine-clinical terms (SNOMED-CT) is the standard lexicon for conditions, procedures and observations. It provides a collection of medical terms with codes for anatomy, diseases, findings, procedures and other domains. For this CPRD CDM transformation, we applied Read to SNOMED-CT mappings provided by the NHS.

We validated this approach by generating the proportion of database records mapped to SNOMED-CT concepts in the CPRD CDM and proportion of terms (Read codes found in the CPRD raw data) mapped to SNOMED-CT concepts for the condition occurrence, procedure occurrence and observation files and reviewed the 100 most frequently occuring conditions, procedures and observations for mapping completeness and accuracy. The 100 most frequent unmapped conditions, procedures and observations were also evaluated to determine if mappings were in fact feasible for these high-frequency codes.

Information loss for conditions was also assessed by comparing condition prevalences for all database years in the CPRD raw data and the CPRD CDM. This was accomplished with a SQL program that estimated prevalence against the raw schema, and an independently written second SQL program that estimated prevalence against the CDM. We examined condition Read codes that occurred during valid patient observation time. To estimate prevalence, we included patients that had an indicator flag for being an 'acceptable' patient and had valid observation time. We considered all condition occurrences where the condition dates fell within the patient's valid observation period and compared the Read code-based prevalence from the raw source with the SNOMED-CT-based prevalence from the CDM. Read codes were analyzed in this manner separately in three groups: those that had a 1:1 mapping with SNOMED-CT concepts, Read codes with the same text description but 'NOS' (not otherwise specified)

grouped in the raw source and conditions where there was more than one Read code such that the Read-to-SNOMED_CT mappings were applied to the CPRD raw data to produce a condition prevalence estimate.

2.1.6 CPRD Lifestyle and Clinical Data

Valuable patient lifestyle information, such as smoking status and body mass index, and clinical measurements, such as blood pressure and laboratory results, are provided in the CPRD data. Because lifestyle and clinical information are potential confounders in observational studies, and laboratory results may be useful for assessing disease status, it was important to include them in the CPRD CDM. CPRD raw lifestyle/clinical data are housed in two tables; within these two tables each data category (e.g. smoking) has a varying number of data elements (e.g. status, cigarettes per day, cigars per day) and these data elements are associated with varying lookups. We created an algorithm to process all data elements in the same manner despite the unusual format described above. Custom source codes were constructed from the data category and data element information and mapped to the Logical Observation Identifiers Names and Codes (LOINC) dictionary concepts (e.g. source code of '4–2' was assigned a source code description of 'Lifestyle Smoking Cigarettes per day') for the implementation.

The algorithms were validated by examining patients with a representative mix of data element types in the raw data against the same patients in the resulting CPRD CDM and by having a second programmer independently code and execute the algorithm and confirm that the results agreed.

2.2 Replication Study Methods

A replication of a prior published study by Schlienger et al. [8] was performed using our instance of raw CPRD data and also the transformed CPRD CDM to compare the results; agreement would serve to further validate the accuracy of our CPRD CDM transformation. Because of changes to the data since the original study was published, it was expected that results found in our raw data study would not have perfect agreement with those reported in the original paper.

2.2.1 Raw Data Analysis

Cases in the Schlienger et al. [8] study had an incident AMI between January 1, 1992 and October 31, 1997. Each patient's observation period began at the latest of: the date the patient's current period of registration with the practice began and the date the practice was deemed to be of research quality, and ended at the earliest of: the date the patient transferred out, the date of last collection of practice data and the patient's date of death; incident AMI diagnoses had to be within the patient's observation period. Patients were required to be aged $\leq$75 years at the date of their AMI (the index date), have an observation period that began at least 3 years prior to that, and not have had one of the following diagnoses between the start of their observation period and 60 days before their index date: AMI, angina pectoris, unexplained chest pain, cardiac arrhythmias, congestive heart failure, stroke, intermittent claudication, venous thromboembolism, chronic renal disease, hypertension, hyperlipidemia, diabetes mellitus, or connective-tissue disorder.

Read code lists for the condition classes referenced above were created using relationships available within the OMOP Standard Vocabularies, all Read code and Multilex code lists for the raw data study mentioned herein are provided in Online Resource 1. Generally, we used the OMOP Standard Vocabularies to generate source code sets for the raw data study rather than the CPRD-provided dictionaries because using the former allows relationships between clinical concepts to be leveraged so that source codes with different terminologies for the same clinical concept can be identified. Standard string searches that can be used instead require a priori knowledge of all possible terminologies. Because some of the condition classes were broad for the prior history exclusion, higher-level SNOMED-CT classification concepts or MedDRA (*Medical Dictionary for Regulatory Activities*) High-level Group or High-level Term concepts were used to extract Read source codes from the OMOP Standard Vocabularies. For example, to gather all cardiovascular disease Read codes for the prior history exclusion, the SNOMED-CT term 'Cardiovascular disease' was used. All cardiovascular disease concepts hierarchically 'below' this concept were identified and Read source codes generated from these concepts. A manual review of these codes was performed to make sure unwanted clinical concepts were not included.

As the original analysis specified, four controls were chosen per case and matched on index date, year of birth, gender, physician practice attended and total observed time prior to the index date. The same exclusions applied to the cases were applied to the controls. As an additional sensitivity analysis, we required controls to exhibit visit activity up to 1 year prior to the index date in addition to the original matching criteria.

NSAID exposures included the following ingredients: acemetacin, diclofenac, diflunisal, etodolac, fenbufen, fenoprofen, flurbiprofen, ibuprofen, indomethacin, ketoprofen, mefenamic acid, nabumetone, naproxen, piroxicam, sulindac, tenoxicam and tiaprofenic acid. Code lists were generated with the OMOP Standard Vocabularies

using the NSAID ingredient concepts above to identify applicable Multilex codes. String searches in the Multilex dictionary by ingredient were also conducted to identify any Multilex codes that may have been missed, e.g. they were not mapped in the OMOP Standard Vocabularies. The NSAID exposures were required to start prior to the patient's index date and within the patient's valid observation period. Patients were defined as a 'current user' if their supply of last NSAID prescription prior to the index date ended at or after the index date, a 'recent user' if their supply ended 1–29 days before the index date, a 'past user' if their supply ended 30 or more days prior to the index date and as non-users if they had no NSAID records prior to the index date. To classify patients as above it was necessary to calculate the duration for each NSAID drug exposure. We used the same duration imputation for the CDM analysis and for the raw data analysis to facilitate comparison. Patients were also classified according to the number of NSAID prescriptions (a proxy for total duration of NSAID therapy) during the patient's valid observation period. 'Current users' were also classified by ingredient.

Potential confounders, body mass index (BMI), smoking status, current aspirin use and long-term hormone replacement therapy (HRT), were assessed manually in the original study from patient profiles. In our analysis, we extracted this information programmatically. BMI (categories: <25, 25–29.9, ≥30 and Unknown) and smoking status (categories: Non, Current, Ex and Unknown) were obtained from the CPRD lifestyle and clinical measurements data in the patient's observation period prior to the index date. Aspirin Multilex codes were generated from the OMOP Standard Vocabularies with an ingredient search. The duration of each aspirin exposure was calculated using the algorithm described above for the CPRD CDM and NSAID exposures. HRT codes were generated from the CPRD drug data dictionary with a BNF (British National Formulary) chapter search (06.04.01.01: Oestrogens and HRT). If the patient had 10 or more HRT prescriptions, she was considered to have been on long-term HRT therapy. A conditional logistic regression model was run for the matched case–control sets to assess AMI risk with the different NSAID categories, adjusted for BMI, smoking, current aspirin use and long-term HRT, and odds ratios (OR) reported with 95 % confidence intervals.

2.2.2 CDM Analysis

We then created a replica of the raw data analysis using the CDM data, including the same case–control analysis to evaluate the risk of first-time AMI associated with NSAID exposure. All source code lists used in the raw data analysis were converted to OMOP concepts using the OMOP Standard Vocabularies; all concept code lists mentioned herein used for the CDM analysis are provided in Online Resource 2. The CDM drug era aggregate file was searched using the NSAID ingredient concepts listed above in the 'NSAID exposures' section of the raw data study. NSAID Multilex codes identified via CPRD dictionary string searches for the raw data analysis that had no mappings and/or valid relationships to ingredients could not be included here as drug exposures in CDM drug eras are collapsed by drug ingredient. The number of NSAID prescriptions per patient were calculated as the total number of prescriptions used to create all NSAID drug eras for the patient. The HRT code list used in the CPRD raw data analysis was converted to ingredients using the OMOP Standard Vocabularies and the drug era file was used to determine HRT exposures. Aspirin exposures were identified using the drug era file and the concept for aspirin.

Patient observation periods were calculated in the CDM with the same algorithm used for the raw data study. Data are not included in the OMOP CDM by definition if they do not occur during the patient's valid observation period. The condition occurrence, procedure occurrence and observation files were searched for conditions instead of the condition era file because the distinction between condition, procedure and observation can often be blurred in the Read dictionary. A procedural record example of this phenomenon is 'Diab mellit insulin–glucose infus acute myocardial infarct' (Read code 889A.00), which was used to identify a prior history of diabetes and AMI for the study exclusion criteria. BMI and smoking data were extracted from the observation file using LOINC BMI and smoking concepts.

3 Results

3.1 CDM Mapping Performance

See Fig. 1 for a comparison of proportions of mapped database records and terms for all domains.

99.9 % of condition records in the condition occurrence file and 98.9 % of condition terms were mapped to the SNOMED-CT dictionary. The top 100 occurring conditions in the CPRD CDM made up 47 % of the condition data and all were mapped and classified correctly (see Table 2 for the top 10 occurring conditions). There were three condition Read codes (out of 46,011) with more than 10,000 records in the data that were unmapped, two had potential SNOMED-CT mappings (N331N00: fragility fracture and 1Z15.00: chronic kidney disease stage 3A) but were not present in the Read to SNOMED-CT mappings provided by the NHS that were used for this transformation.

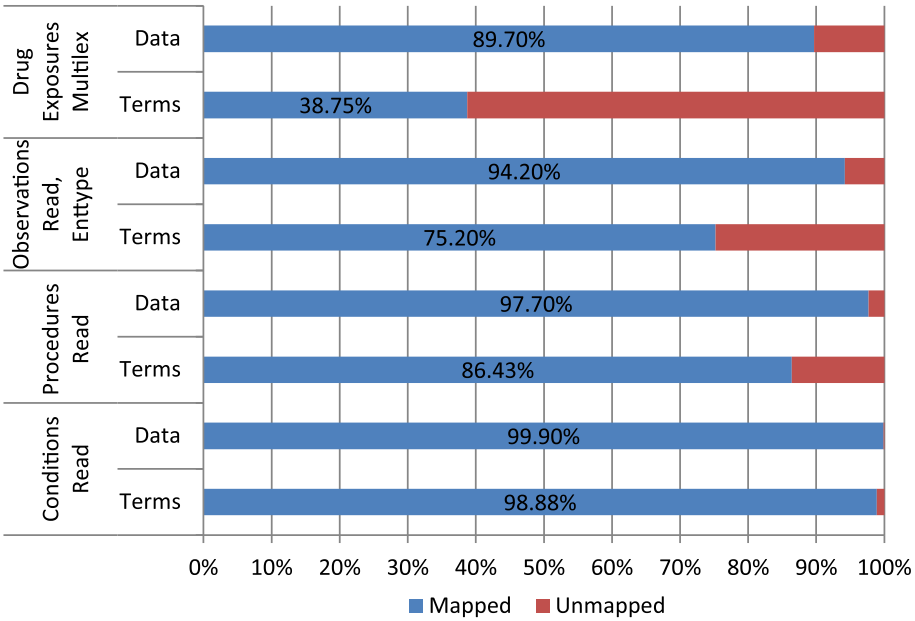

Fig. 1 Proportion of mapped terms and database records for the CPRD OMOP CDM domains. *CDM* common data model, *CPRD* clinical practice research datalink, *OMOP* observational medical outcomes partnership

Table 2 Top 10 conditions and procedures with mapped concept description and prevalence in the CPRD OMOP CDM

Domain	Source code/description	SNOMED-CT	Prevalence (patients with source code/all patients in database)
Condition	H05z.00 Upper respiratory infection NOS	Acute upper respiratory infection	12.3
Condition	H05z.11 Upper respiratory tract infection NOS	Acute upper respiratory infection	9.6
Condition	H03..00 Acute tonsillitis	Acute tonsillitis	8.2
Condition	H06z011 Chest infection	Lower respiratory tract infection	8.2
Condition	H06z000 Chest infection NOS	Lower respiratory tract infection	6.9
Condition	N131.00 Cervicalgia: pain in neck	Nonspecific pain in the neck region	6.7
Condition	N142.11 Low back pain	Low back pain	6.6
Condition	N245.17 Shoulder pain	Shoulder pain	6.2
Condition	H33..00 Asthma	Hyperreactive airway disease	5.7
Condition	G20..00 Essential hypertension	Essential hypertension	5.5
Procedure	535..00 Standard chest X-ray	Plain chest X-ray	8.4
Procedure	3395.00 Peak expiratory flow rate: PEFR/PFR	PEFR: peak expiratory flow rate	6.7
Procedure	7305011 Syringe ear to remove wax	Syringing ear to remove wax	5.9
Procedure	81H..00 Dressing of wound	Dressing of wound	5.6
Procedure	6637.00 Inhaler technique observed	Inhaler technique observed	4.8
Procedure	662..12 Hypertension monitoring	Hypertension monitoring	4.5
Procedure	8C1B.00 Nursing care blood sample taken	Blood sample taken	4.3
Procedure	7L17200 Blood withdrawal for testing	Taking blood sample	4
Procedure	7L17.00 Blood withdrawal	Phlebotomy	3.7
Procedure	663..11 Asthma monitoring	Asthma monitoring	3.3

CDM common data model, *CPRD* clinical practice research datalink, *OMOP* observational medical outcomes partnership

Condition prevalences were compared across the CPRD CDM and the CPRD raw data to assess information loss. 15,130 condition Read codes out of the 32,758 mapped (46.2 %) had a 1:1 mapping with SNOMED-CT concepts. All prevalences in the raw data associated with these Read codes were exactly equivalent to the CPRD CDM prevalences from the corresponding SNOMED_CT codes. We also grouped Read codes with the same text description but 'NOS' (not otherwise specified) indicated for one of the two that were mapped to the same SNOMED-CT concept.

Table 3 Top 10 observations and drug exposures with mapped concept description and prevalence in the CPRD OMOP CDM

Domain	Source code/description	SNOMED/LOINC concept description	Prevalence (patients with source code/all patients in database)
Observation	Lifestyle smoking status smoke/drink status	PhenX measure: tobacco: smoking status	64.5
Observation	Examination findings blood pressure systolic	Diastolic blood pressure	60.3
Observation	Examination findings blood pressure diastolic	Systolic blood pressure	60.3
Observation	Examination findings weight in kilograms	Body weight measured	59.6
Observation	22A..00 O/E: weight	O/E: weight	59.5
Observation	246..00 O/E: blood pressure reading	O/E: blood pressure reading	58.4
Observation	Examination findings weight BMI	Body mass index (BMI) [ratio]	50.7
Observation	9N31.00 Telephone encounter	Telephone encounter	38.5
Observation	Serum creatinine	Creatinine [mass/volume] in serum or plasma	35.9
Observation	44J3.00 Serum creatinine	Creatinine measurement, serum	35.2
Drug exposure	FLU data not entered 65E.00 influenza vaccination	Influenza virus vaccine	19.4
Drug exposure	58932020 Paracetamol 500-mg tablets	Acetaminophen 500-mg oral tablet	11.3
Drug exposure	69782020 Omeprazole 20-mg gastro-resistant capsules	Omeprazole 20-mg enteric-coated capsule	10.4
Drug exposure	52994020 Aspirin 75-mg dispersible tablets	Aspirin 75-mg disintegrating tablet	7.1
Drug exposure	58976020 Bendroflumethiazide 2.5-mg tablets	Bendroflumethiazide 2.5-mg oral tablet	6.3
Drug exposure	72489020 Simvastatin 40-mg tablets	Simvastatin 40-mg oral tablet	5.2
Drug exposure	59420020 Furosemide 40-mg tablets	Furosemide 40-mg oral tablet	3.9
Drug exposure	72488020 Simvastatin 20-mg tablet	Simvastatin 20-mg oral tablet	3.7
Drug exposure	60153020 Atenolol 50-mg tablet	Atenolol 50-mg oral tablet	3.2
Drug Exposure	55991020 Levothyroxine sodium 100-μg tablet	Levothyroxine sodium 1-mg oral tablet	1.9

CDM common data model, *CPRD* clinical practice research datalink, *OMOP* observational medical outcomes partnership

This group contained 1,282 Read codes and made up 3.9 % of the mapped condition codes. All prevalences in the raw data associated with these Read codes were also exactly equivalent to the CPRD CDM prevalences from the corresponding SNOMED_CT codes. The remaining Read codes were mapped to SNOMED-CT codes in the raw data and raw data prevalences and CPRD CDM prevalences were identical.

97.7 % of procedure records in the procedure occurrence file and 86.4 % of procedure terms were mapped to the SNOMED-CT dictionary. The top 100 occurring procedures in the CPRD CDM made up 68.5 % of the procedure data and all were mapped and classified correctly (Table 2 for the top 10 occurring procedures). Among 23,278 procedure Read codes with more than 10,000 records, 31 were unmapped. Most of these were patient management codes such as '66QC.00: anticoagulation monitoring—secondary care' and 'Z174N00: wound care'.

The final observation file with all three observation record sources (lifestyle/clinical, laboratory and Read code observation data) had 94.2 % of observations records and 75.2 % of observation terms mapped to either the LOINC or SNOMED-CT dictionaries. The top 100 occurring observations in the CPRD CDM made up 60 % of the observation data and all were mapped and classified correctly except four unmappable codes from the lifestyle/clinical data (Table 3 for the top 10 occurring observations). Most unmapped observation codes with greater than 10,000 records in the data could not have been mapped. For example 'Lifestyle Contraception Date IUCD fitted' from the lifestyle/clinical file could not be mapped to the LOINC dictionary and '9D1..00 MED3—doctor's statement' could not be mapped to the SNOMED-CT dictionary.

89.7 % of drug records and 38.8 % of drug terms in the CPRD CDM were mapped to drug concepts. 2.1 % of these mapped drug records did not have relationships to RxNorm ingredients in the OMOP Standard Vocabularies. The top 100 drug exposures in the CPRD CDM make up 42 % of the drug exposure data and all were mapped correctly except six products (five over the counter (OTC)) not available in the US such as 'Adcal-D3 chewable tablets tutti frutti (ProStrakan Ltd)' (Table 3 for the top 10 occurring drugs). The majority of unmapped drug exposures with more than 10,000 records represented in the data were medical devices/supplies and OTC products such as 'Dermol cream (Dermal Laboratories Ltd)' and 'U100 Insulin syringe 0.5 ml' with a few UK products not available in the US such as 'Seretide 500 Accuhaler (GlaxoSmithKline UK Ltd)'.

Drug prevalences for RxNorm concepts for all database years between the CPRD raw data (drug exposures mapped

to RxNorm) and the CPRD CDM were compared for 21,533 Multilex codes that occurred during valid patient observation time. In this analysis, 99.80 % of these codes had equivalent prevalences when the CPRD CDM and the CPRD raw data were compared.

Seven percent of drug records had valid duration values in the raw drug exposure data. Durations were imputed for 89.3 % of drug exposure records using the most common duration for combinations of product, quantity, numeric daily dose and numpacks (Step 1 in duration imputation process). The remaining duration imputations were calculated using the most common duration value for product only, or were assumed to have been 1 day (Steps 2 and 3 in the duration imputation process). After filtering out valid imputations as defined in the methods section that were calculated in Steps 1–3, 228,437 imputed duration values for combinations of product, quantity, numeric daily dose and numpacks remained out of a possible 1,923,613 (11.9 % of imputations representing 3.1 % of all drug exposure data) that were possibly erroneous. Of these imputation combinations, only 595 had greater than 10,000 drug exposure records in the data. The remaining was generally inferred from a small number of records, and was applied to a small number of records. We reviewed the most frequently occurring possibly erroneous imputations and found that very few were oral prescription drugs for which the notion of continuous exposure is more straightforward; instead the majority of these products were inhalers, creams, nasal sprays, and ear drops where inferring duration is more difficult given the source data. For all database records with a valid numeric daily dose (>0), 68.1 % of database records with imputed duration were equivalent to quantity/numeric daily dose and 74.7 % of database records had an absolute difference of no greater than 5 between quantity/numeric daily dose and imputed duration.

3.2 Replication Study Results

Table 4 presents characteristics of cases for the original, raw data and CDM studies. 3,315 cases and 13,139 controls were analysed in the original study. In our raw data replication of the original study, 3,458 cases were found and 13,165 controls were matched. In our CDM replication of the original study, 3,454 cases were found and 13,159 controls were matched. In the sensitivity study, 3,458 cases were found and 12,891 controls were matched. Age distributions for cases in the raw data study vs. the original study were inconsistent, with more patients in the 70–75 years age group in the raw data study and more patients in the age <40 years group in the original study. Age distributions for cases in the raw data study were the same as in

Table 4 Characteristics of cases and controls for replication study

	Cases			Controls			
	Original study ($N = 3,315$) *N* (%)	Raw data study ($N = 3,458$) *N* (%)	CDM study ($N = 3,454$) *N* (%)	Original study ($N = 13,139$) *N* (%)	Raw data study ($N = 13,165$) *N* (%)	CDM study ($N = 13,159$) *N* (%)	Sensitivity study ($N = 12,891$) *N* (%)
Age (years)							
<40	91 (2.8)	67 (1.9)	67 (1.9)	367 (2.8)	259 (2.0)	260 (2.0)	252 (2.0)
40–49	417 (12.6)	373 (10.8)	373 (10.8)	1,656 (12.6)	1,441 (10.9)	1,443 (11.0)	1,423 (11.0)
50–59	830 (25.0)	876 (25.3)	875 (25.3)	3,314 (25.2)	3,330 (25.3)	3,326 (25.3)	3,276 (25.4)
60–69	1,227 (37.0)	1,263 (36.5)	1,261 (36.5)	4,832 (36.8)	4,836 (36.7)	4,830 (36.7)	4,732 (36.7)
70–75	750 (22.6)	879 (25.4)	878 (25.4)	2,970 (22.6)	3,304 (25.1)	3,300 (25.1)	3,208 (24.9)
Male	2,452 (74.0)	2,546 (73.6)	2,543 (73.6)	9,715 (73.9)	9,692 (73.6)	9,685 (73.6)	9,463 (73.4)
Female	863 (26.0)	912 (26.4)	911 (26.4)	3,424 (26.1)	3,478 (26.4)	3,474 (26.4)	3,428 (26.6)
Smoking status							
Non	1,079 (32.6)	924 (26.7)	921 (26.7)	6,204 (47.2)	4,475 (34.0)	4,510 (34.3)	5,157 (40.0)
Current	1,100 (33.2)	963 (27.8)	964 (27.9)	2,574 (19.6)	2,074 (15.7)	2,112 (16.0)	2,286 (17.7)
Ex	376 (11.3)	291 (8.4)	290 (8.4)	1,353 (10.3)	904 (6.9)	888 (6.7)	1,069 (8.3)
Unknown	760 (22.9)	1,280 (37.0)	1,279 (37.0)	3,008 (22.9)	5,717 (43.4)	5,649 (42.9)	4,379 (34.0)
Body mass index (kg/m^2)							
<25	885 (26.7)	740 (21.4)	738 (21.4)	4,240 (32.3)	2,753 (20.9)	2,861 (21.7)	3,216 (24.9)
25–29	1,100 (33.2)	798 (23.1)	798 (23.1)	4,004 (30.5)	2,874 (21.8)	2,892 (22.0)	3,275 (25.4)
≥30	387 (11.7)	318 (9.2)	318 (9.2)	1,208 (9.2)	905 (6.9)	880 (6.7)	1,021 (7.9)
Unknown	943 (28.4)	1,602 (46.3)	1,600 (46.3)	3,687 (28.0)	6,638 (50.4)	6,526 (49.6)	5,379 (41.7)

Table 5 Risk of developing first-time acute myocardial infarction adjusted for body mass index, smoking status, aspirin use and hormone replacement therapy

	Original study adjusted odds ratio	Raw data study adjusted odds ratio	CDM study adjusted odds ratio	Sensitivity study adjusted odds ratio
Smoking status				
Current	2.7 (2.4–2.9)	2.3 (2.0–2.6)	2.4 (2.1–2.7)	2.4 (2.1–2.7)
Ex	1.6 (1.4–1.9)	1.6 (1.3–1.9)	1.5 (1.3–1.9)	1.5 (1.3–1.9)
Unknown	1.5 (1.3–1.7)	1.2 (1.0–1.3)	1.2 (1.0–1.4)	1.6 (1.4–1.8)
Body mass index (kg/m^2)				
25–29.9	1.4 (1.3–1.6)	1.1 (1.0–1.3)	1.1 (1.0–1.3)	1.2 (1.0–1.3)
≥30	1.7 (1.4–1.9)	1.5 (1.2–1.9)	1.6 (1.2–2.0)	1.6 (1.3–2.0)
Unknown	1.2 (1.1–1.4)	1.1 (1.0–1.3)	1.2 (1.0–1.4)	1.3 (1.1–1.4)

Table 6 Current, recent past and past exposure to NSAIDs, stratified by duration of NSAID therapy (number of prescriptions) and ingredient exposure in current users

NSAID category/ number of prescriptions	Cases			Controls			
	Original study (N = 3,315) N (%)	Raw data study (N = 3,458) N (%)	CDM study (N = 3,454) N (%)	Original study (N = 13,139) N (%)	Raw data study (N = 13,165) N (%)	CDM study (N = 13,159) N (%)	Sensitivity study (N = 12,891) N (%)
Non-users	1,502 (45.3)	1,966 (56.9)	1,981 (57.4)	6,236 (47.5)	8,260 (62.7)	8,269 (62.8)	7,099 (55.1)
Current	242 (7.3)	206 (6.0)	187 (5.4)	825 (6.3)	560 (4.3)	584 (4.4)	683 (5.3)
Current 1–4	34 (1)	42 (1.2)	41 (1.2)	111 (0.8)	121 (0.9)	129 (1.0)	175 (1.4)
Current 5–9	45 (1.4)	30 (0.9)	27 (0.8)	157 (1.2)	83 (0.6)	85 (0.6)	105 (0.8)
Current 10–19	36 (1.1)	34 (1.0)	31 (0.9)	145 (1.1)	120 (0.9)	98 (0.7)	140 (1.1)
Current 20–29	38 (1.1)	36 (1.0)	29 (0.8)	119 (0.9)	88 (0.7)	86 (0.7)	97 (0.8)
Current 30+	89 (2.7)	64 (1.9)	59 (1.7)	293 (2.2)	148 (1.1)	186 (1.4)	166 (1.3)
Recent past	118 (3.6)	116 (3.4)	105 (3.0)	377 (2.9)	330 (2.5)	258 (2.0)	389 (3.0)
Recent past 1–4	25 (0.8)	37 (1.1)	36 (1.0)	105 (0.8)	126 (1.0)	134 (1.0)	173 (1.3)
Recent past 5–9	21 (0.6)	16 (0.5)	14 (0.4)	95 (0.7)	72 (0.5)	48 (0.4)	74 (0.6)
Recent past 10–19	23 (0.7)	24 (0.7)	23 (0.7)	77 (0.6)	65 (0.5)	42 (0.3)	57 (0.4)
Recent past 20–29	14 (0.4)	16 (0.5)	15 (0.4)	44 (0.3)	24 (0.2)	16 (0.1)	44 (0.3)
Recent past 30+	35 (1.1)	23 (0.7)	17 (0.5)	56 (0.4)	43 (0.3)	18 (0.1)	41 (0.3)
Past	1,453 (43.8)	1,170 (33.8)	1,181 (34.2)	5,701 (43.4)	4,020 (30.5)	4,048 (30.8)	4,720 (36.6)
Past 1–4	984 (29.7)	915 (26.5)	921 (26.7)	4,002 (30.5)	3,242 (24.6)	3,275 (24.9)	3,753 (29.1)
Past 5–9	311 (9.4)	136 (3.9)	132 (3.8)	1,190 (9.1)	482 (3.7)	449 (3.4)	607 (4.7)
Past 10–19	91 (2.7)	77 (2.2)	76 (2.2)	352 (2.7)	193 (1.5)	219 (1.7)	243 (1.9)
Past 20–29	26 (0.8)	25 (0.7)	24 (0.7)	82 (0.6)	61 (0.5)	55 (0.4)	60 (0.5)
Past 30+	41 (1.2)	17 (0.5)	28 (0.8)	75 (0.6)	42 (0.3)	50 (0.4)	57 (0.4)
Ibuprofen	60 (24.8)	68 (33.0)	64 (34.2)	204 (24.7)	144 (25.7)	173 (29.6)	189 (27.7)
Diclofenac	97 (40.1)	68 (33.0)	62 (33.1)	277 (33.6)	173 (30.9)	179 (30.7)	214 (31.3)
Piroxicam	10 (4.1)	9 (4.4)	9 (4.8)	28 (3.4)	34 (6.1)	41 (7.0)	33 (4.8)
Ketoprofen	15 (6.2)	5 (2.4)	3 (1.6)	48 (5.8)	32 (5.7)	26 (4.5)	35 (5.1)
Indomethacin	15 (6.2)	21 (10.0)	20 (10.7)	56 (6.8)	44 (7.9)	58 (9.9)	55 (8.1)
Naproxen	19 (7.9)	14 (6.8)	13 (7.0)	105 (12.7)	69 (12.3)	58 (9.9)	78 (11.4)

the CDM study. Gender distributions were similar in the raw data study compared with the original study: 74 % cases were male in the original study, 73.6 % in the raw data study and in the CDM study. BMI was unknown for 46.3 % of cases in the raw data study and the CDM study vs. 28.4 % of cases in the original study. Smoking status was unknown for 37 % of cases in the raw data study and the CDM study vs. 22.9 % of cases in the original study.

All three studies identified an increased risk of developing AMI for 'Current' smoking groups vs. Non-smokers (Table 5) and the estimates were similar. The original study found an adjusted OR of 2.7, with a 95 % confidence

interval (CI) between 2.4 and 2.9. The raw data study found an OR of 2.3 (95 % CI 2.0–2.6), the CDM study 2.4 (95 % CI 2.1–2.7). Similarly, all three found increased and similar risks of AMI for BMI >30 vs. <25. The original study had an adjusted OR of 1.7 (95 % CI 1.4–1.9), the raw data study 1.5 (95 % CI 1.2–1.9) and the CDM study 1.6 (95 % CI 1.2–2.0).

Table 6 shows, by NSAID exposure status, case counts for the original, raw data, and CDM studies; and control counts for these studies and the sensitivity analysis. 'Current', 'Recent past' and 'Past' exposure counts along with counts stratified by duration of treatment using number of NSAID prescriptions as a proxy are presented. In addition, NSAID ingredient counts are calculated for the 'Current' user group. The percentage of NSAID non-users for cases is greater in the raw data study vs. the original study (56.9 vs. 45.3 % respectively). The percentage of control group NSAID non-users in the sensitivity analysis is lower than in the raw data study (55.1 vs. 62.7 % respectively). For almost all categories for cases, fewer NSAID exposures were found in the raw data study vs. the original study with the 'Past' user category for cases having the biggest relative difference (33.8 vs. 43.8 % of cases). The above is even more pronounced in a control group comparison with the original study finding 43.4 % of controls with NSAID exposures in the 'Past' user category, and the raw data study finding 30.5 %. 36.6 % of controls in the sensitivity analysis had NSAID exposures in the 'Past' user category.

The percentage of NSAID case non-users in the CDM study is slightly higher than in the raw data study (57.4 vs. 56.9 %, respectively). The NSAID code list used in the studies was analysed and 85 out of 861 NSAID codes had no relationship to an ingredient in the OMOP Standard Vocabularies, either because they were unmapped or had broken relationships. These 85 codes represent 3 % of NSAID records in the CPRD database represented by the 861 identified NSAID codes. For almost all categories for cases excepting the 'Past NSAIDs' category, slightly fewer NSAID exposures were found in the CDM study vs. the raw data study. Thirty-three percent of 'Current NSAID' users had ibuprofen as their last NSAID prior to index date in the raw data analysis, vs. 24.8 % in the original study, and 34.2 % in the CDM study. 33 % of 'Current NSAID users' had diclofenac as their last NSAID prior to index

Table 7 Risk of first-time acute myocardial infarction associated with current, recent past and past NSAID therapy, duration and ingredient in current users

NSAID category/number prescriptions	Original study adjusted odds ratio	Raw data study adjusted odds ratio	CDM study adjusted odds ratio	Sensitivity study adjusted odds ratio
Current	1.17 (0.99–1.37)	1.51 (1.24–1.83)	1.29 (1.06–1.57)	1.07 (0.89– 1.29)
Current 1–4	1.30 (0.87–1.93)	1.32 (0.87–2.01)	1.15 (0.77– 1.74)	0.80 (0.54– 1.18)
Current 5–9	1.10 (0.78–1.54)	1.44 (0.87–2.40)	1.32 (0.79– 2.21)	1.03 (0.64– 1.67)
Current 10–19	0.97 (0.66–1.42)	1.40 (0.89–2.19)	1.31 (0.83– 2.07)	1.05 (0.69– 1.62)
Current 20–29	1.31 (0.89–1.91)	1.58 (0.99–2.54)	1.45 (0.88–2.40)	1.17 (0.75–1.82)
Current 30+	1.21 (0.94–1.55)	1.71 (1.20–2.45)	1.30 (0.91– 1.85)	1.28 (0.90– 1.83)
Recent past	1.26 (1.01–1.57)	1.41 (1.09–1.82)	1.63 (1.24– 2.16)	1.20 (0.94– 1.55)
Recent past 1–4	0.95 (0.61–1.48)	1.24 (0.80–1.93)	1.21 (0.79– 1.87)	0.77 (0.51– 1.16)
Recent past 5–9	0.90 (0.55–1.46)	0.95 (0.51–1.78)	1.26 (0.60– 2.68)	1.10 (0.56– 2.18)
Recent past 10–19	1.13 (0.70–1.83)	1.36 (0.77–2.41)	1.66 (0.91–3.03)	1.88 (1.05–3.37)
Recent past 20–29	1.33 (0.72–2.46)	2.35 (1.11–5.01)	3.79 (1.61–8.90)	1.65 (0.82, 3.32)
Recent past 30+	2.71 (1.75–4.22)	2.09 (1.12–3.91)	3.25 (1.43–7.36)	2.00 (1.07–3.73)
Past	1.04 (0.96–1.13)	1.21 (1.10–1.32)	1.17 (1.07–1.28)	0.89 (0.82–0.97)
Past 1–4	1.02 (0.93–1.12)	1.18 (1.07–1.30)	1.14 (1.04–1.26)	0.89 (0.81–0.98)
Past 5–9	1.06 (0.92–1.22)	1.07 (0.85–1.35)	1.07 (0.84–1.35)	0.75 (0.59–0.94)
Past 10–19	1.00 (0.78–1.28)	1.82 (1.31–2.53)	1.26 (0.92–1.74)	1.15 (0.84–1.58)
Past 20–29	1.26 (0.80–2.01)	1.13 (0.62–2.06)	1.73 (0.94– 3.19)	1.42 (0.77–2.61)
Past 30+	2.33 (1.57–3.46)	1.57 (0.81–3.02)	2.44 (1.34–4.45)	1.07 (0.55–2.07)
Ibuprofen	1.17 (0.87–1.58)	1.73 (1.23–2.44)	1.47 (1.04–2.08)	1.16 (0.84–1.61)
Diclofenac	1.38 (1.08–1.77)	1.65 (1.17–2.33)	1.53 (1.08–2.16)	1.11 (0.80–1.53)
Piroxicam	1.65 (0.78–3.49)	0.93 (0.38– 2.24)	0.76 (0.33–1.72)	0.73 (0.29–1.80)
Ketoprofen	1.39 (0.77–2.51)	0.64 (0.20– 1.99)	0.59 (0.16–2.19)	0.55 (0.16–1.95)
Indomethacin	1.03 (0.58–1.85)	1.96 (1.04– 3.72)	1.29 (0.70–2.37)	1.68 (0.91–3.09)
Naproxen	0.68 (0.42–1.13)	0.81 (0.41– 1.61)	0.69 (0.35–1.37)	0.59 (0.31–1.14)

date in the raw data analysis, vs. 40.1 % in the original study, and 33.1 % in the CDM study (Table 6).

Duration imputations used for the raw data and the CDM studies were assessed by comparing distributions of current, recent, and past users in the raw data study to the original study. Proportions of NSAID users for cases and controls for current, recent past and past users were distributed fairly equally (with past users containing most of the extra exposures not found in the raw data study). For the original study, current NSAID users were 7.3 % of cases, recent past NSAID users were 3.6 % of cases and past NSAID users were 43.8 % of users. In the raw data study, current NSAID users were 6.0 % of cases, recent past NSAID users were 3.4 % of cases and past NSAID users were 33.8 % of users (Table 6).

For the majority of NSAID exposure categories, point estimates of developing an AMI with NSAID therapy were higher in the raw data study than the original, similar in the CDM vs. raw data studies and similar in the sensitivity analysis vs. the original study. The adjusted odds ratio for the 'Current' user NSAID group was 1.17 (95 % CI 0.99–1.37) in the original study, 1.51 (95 % CI 1.24–1.83) in the raw data study, 1.29 (95 % CI 1.06–1.57) in the CDM study and 1.07 (95 % CI 0.89–1.29) in the sensitivity analysis. The original and CDM studies showed increased risk of developing AMI with NSAID therapy in 'Recent past' and 'Past' user groups with 20–29 and 30+ prescriptions; this trend of increased risk based on treatment duration materialized in our raw data and sensitivity studies in all user groups (Table 7).

4 Discussion

4.1 Data Transformation and Source Code Mapping

CPRD can be accurately transformed into the OMOP CDM, with acceptable database information loss with respect to drug, condition and observation records; 10.3 % of drug exposures, 0.15 % of conditions, 2.3 % of procedures and 5.8 % of observations were unmapped. Drug exposure loss appears higher than desired; however, the majority of unmapped drug exposures were OTC products and medical devices/supplies, while most procedure and observation data that were unmapped included patient management records. Matching of CPRD source codes to semantically identical OMOP concepts and domain classification efforts were also assessed and found to be of high quality via manual review of the top 100 mapped and unmapped source codes for a domain. Drug and condition prevalence comparisons between selected mapped terms in the CPRD CDM and the CPRD raw data were equivalent.

The Zhou et al. [7] transformation of the THIN database yielded unmapped record proportions of 7 % for drug exposures, 6 % for conditions and 4 % for procedures in their resulting CDM. They concluded the extent of information loss from unmapped records and the challenging data structure of THIN limited their CDM's usefulness for pharmacoepidemiological research, primarily because of database drug exposure and condition loss. Based on our evaluation of the available condition, drug, and procedure code mapping, the findings from the comparison of condition and drug prevalence between the source and CDM, and the consistent results from our replication study, we believe that the transformed CPRD CDM is suitable for use in observational research. Though it appeared we had more drug exposure information loss than Zhou et al. [7], most of the drug exposure records that could not be mapped may not be useful for drug safety surveillance and epidemiological analyses and our replication study results indicated that drug loss information was not enough to overly affect the results of the CDM study vs. the raw data study. In addition, all CPRD data available were converted to the CPRD CDM for this transformation, including the entire body of lifestyle/clinical and laboratory data, despite its unusual data structure. These data were transformed systematically with the same algorithm handling all test types.

It is important to reinforce that while the CDM provides the opportunity to normalize all codes into a common reference standard that is applied consistently across all databases, the CDM also maintains the source codes from the original raw database. As a result, while the CDM makes it efficient to do cross-database analyses under a standard vocabulary, if a specific research question requires analysis with the local source codes (Read conditions and Multilex drugs for CPRD), then that is fully supported.

In this version of the CPRD CDM, we chose to exclude medical history information that fell prior to the valid observation period to adhere to CDM convention. However, it is conceivable that medical history records, though potentially incomplete in CPRD, can be relevant information depending upon the design of the observational study in question. In future CPRD CDM versions, we will be exploring methods to include these data without deviating from CDM model requirements.

Our duration imputation validation performed by the use case and comparisons of quantity/numeric daily dose and the duration imputation had acceptable results, but because 26 % of CPRD source drug exposures have no daily dose information, we believe source record verification would need to be evaluated to create a best practice for inference. While CPRD requires extensive imputation to infer length of drug exposure, other data sources may require different approaches. Further research across a data network should

assess the impact of imputation algorithms on effect estimates.

Transformation of a source database to a common data model requires a diverse set of skills, including expertise concerning both the source data and the destination model. Based on our experience, we believe the best practice in establishing a transformation involves close collaboration between the data holder and the research data network architects. In our case, the Janssen team that licenses the CPRD data worked directly with the CPRD staff to address questions about the source data, and also worked directly with the OMOP team to ensure that the application of the CDM standard and vocabulary was consistent with other community efforts. Within a research data network it is important to apply conventions consistently across all databases, but it is essential that those knowledgeable about the source data provide their expertise to ensure the conventions are applied correctly within their source. It is important to reinforce that a standardized transformation of a source database does not remove the need for researchers to have a complete understanding of the underlying data. The CDM provides a consistent structure and allows for a standardized data quality process to be applied within the analysis framework, but researchers still require domain expertise and empirical evidence to support the use of any data source for any specific research question.

4.2 Raw Data Analysis vs. CDM Analysis

How useful a particular CDM will be for epidemiologic analyses can be separated into two different concepts: adequacy of the CDM structure (can the model accommodate the variables you need for the analysis) and the content (are the values for the variables you need faithfully captured). We believe with our particular use case (replication of original study examining risk of AMI with NSAIDs) we have shown CDM structure meets the need, and though source code mappings can be improved, present mappings did not overly perturb our analysis.

Only 0.12 % of cases that were found in the raw data study were not found in the CDM study. This validates the fidelity of the CDM condition Read data transformation for a wide array of conditions because of extensive prior history condition exclusions. Patient characteristics in both studies were the same for cases, validating that demographic (person file) and lifestyle data (observation file) information were transformed accurately. Slightly fewer NSAID exposures were found in the CDM cases vs. raw data cases because of unmapped NSAID codes and NSAID codes with broken relationships in the OMOP Standard Vocabularies. Unmapped NSAID drug exposures occurring later in a drug era caused a small number of cases in the CDM study to be classified as 'Past NSAID' users instead of 'Current NSAID' or 'Recent past NSAID' users. The proportion of unmapped NSAID codes and NSAID codes lacking relationships to ingredients would most likely be lower for studies conducted with more recent data as the legacy codes identified above generally are not mapped in the OMOP Standard Vocabularies, and appear rarely in current data. After stratification on use type ('Current', 'Recent past' and 'Past'), the ORs in these two studies were similar. This validates the drug exposure data transformation (drug exposure and drug era), as the small number of unmapped drugs did not overly affect the final ORs. The variability shown in ORs with categories stratified by duration (especially in the 30+ prescription categories) was probably the result of smaller cell sizes.

The CDM replication analysis was easier to perform and required much less programmatic effort than the raw data study (8 h total programming time for the CDM analysis vs. 40 h for the raw data analysis) owing to the standardized structure of the data, vocabulary queries that leverage relationships between concepts and useful derived constructs such as the drug era file. In addition, the quality of additional analyses will be improved as validated algorithms within the CDM can be leveraged. In other words, certain methodological decisions (e.g. calculating patient observation periods) are only made once during the CDM transformation process and need not be re-addressed with each subsequent CDM analysis performed. The CPRD CDM can also be a valuable part of future efforts to compare CPRD with other observational databases.

4.3 Raw Data Analysis vs. Original Study

The objective of this effort was to determine if the CDM version of the CPRD data was a good approximation of the raw data; however the comparison also indicated there were differences between the CPRD data from the time period of the original study vs. the data as it exists today. The smaller number of NSAID exposures found in the raw data study for cases and controls vs. the original study may reflect: the BCDSP requirement that a patient have a prescription (of any drug, not necessarily an NSAID) at least 3 years prior to the AMI to guarantee that the patient was active in the database, there may have been errors in the older drug record conversions to Multilex that rendered some older drug records unavailable in current CPRD versions and differing practice composition between the older and newer versions of CPRD with newer practices having less historical data. Larger proportions of patients with 'Unknown' BMI and smoking status in the raw data study are most likely the result of: contrasting observation period algorithms and differing practice composition between the two data cuts as explained above resulting in less historical data present in newer practices. We believe

more first-time AMI patients were found in the raw data study vs. the original study because practices that do not meet BCDSP quality standards are retroactively eliminated from their CPRD database [3].

The fact that proportions of NSAID users for cases and controls for current, recent past and past users in the original study vs. the raw data study were distributed fairly equally and validated the CPRD CDM duration imputation against a published study.

ORs from the raw data analysis were somewhat higher than those from the original study and existing literature. Across the board fewer NSAID exposures were identified in controls from the raw data study than in the original study so it appeared there was a difference in control group selection between the two. After completion of our raw data study, the BCDSP shared with us the algorithm they used for defining patient observation periods. The first prescription date of any drug was used as the start of the patient's observation period for cases and controls (controls were also required to have a visit or prescription any time after the case's index date). Hence, controls without prescriptions at least 3 years prior to the case's index date were not eligible for the study (3 years observation time required prior to index date), forcing controls to be active in the database in the original study for at least 3 years prior to the case's index date. In the control matching algorithm for the raw data analysis we did not require healthcare use by patients prior to the case's index date, so healthier patients were not excluded as they were in the original analysis.

To test the hypothesis that requiring evidence of healthcare activity by matching controls on the visit date prior to the case's index date would increase NSAID exposures in the controls and lower the OR, the sensitivity data study was performed [13]. Adjusted ORs very similar to the original study, and lower than those from the the raw data analysis or the CDM analysis were observed for AMI risk with NSAIDS. As expected, controls had more NSAID exposures in the sensitivity data analysis vs. the raw data analysis. Thus, matching controls on a visit at approximately the case's index date appears to be a useful strategy to combat what would otherwise be a positive bias in case–control studies.

5 Conclusions

Our research leads us to the belief that CPRD can be accurately transformed into the OMOP CDM with acceptable information loss across drugs, conditions and observations. Matching of CPRD source codes to semantically identical OMOP concepts and domain classification efforts were assessed to be of high quality through manual review of high-frequency source codes. In addition, we determined that for a particular use case (risk of AMI with NSAIDs) the CDM structure was adequate and the mappings could be improved but did not overly perturb our analysis, and drug and condition prevalences between the CPRD raw data and the CPRD CDM were comparable. The drug exposure duration imputation required for the CDM also compared favourably with the use case. The CPRD CDM replication analysis required much less programmatic effort than the raw data study and quality of additional analyses will be improved as validated algorithms within the CDM can be leveraged. Additionally, the CPRD CDM can be a valuable part of future efforts to compare CPRD with other observational databases.

Acknowledgments We thank Dr. Susan Jick and Dr. Hershel Jick for their assistance in understanding changes in the CPRD database over time and key details of the methods that may have contributed to differences between the original and raw data studies, John Logie and Lakshmi Mynepalli from GlaxoSmithKline for sharing the duration imputation method used in their CPRD to CDM transformation with the OMOP community that our imputation was based upon, Anton Ivanov and Frank DeFalco for developing CDMBuilder, Erica Voss for managing the Standard Vocabularies at Janssen and for review, and Jesse Berlin and Rupa Makadia for review. This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in this study are those of the authors alone. The protocol for this study (reference # 14_076) is provided with the submission and was approved by the Independent Scientific Advisory Committee. No sources of funding were used to conduct this study or prepare this manuscript. Amy Matcho, Patrick Ryan and Daniel Fife are full-time employees of Janssen Research and Development, LLC and shareholders of Johnson and Johnson. Christian Reich is a full-time employee of AstraZeneca PLC. Both companies market ibuprofen, a NSAID.

References

1. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the general practice research database and linked cancer registries. Pharmacoepidemiol Drug Saf. 2013;22(2):168–75. doi:10.1002/pds.3374.
2. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the general practice research database: a systematic review. Br J Clin Pharmacol. 2010;69(1):4–14. doi:10.1111/j.1365-2125.2009.03537.x.
3. Jick SS, Kaye JA, Vasilakis-Scaramozza C, Garcia Rodriguez LA, Ruigomez A, Meier CR, et al. Validity of the general practice research database. Pharmacotherapy. 2003;23(5):686–9.
4. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med. 2010;153(9):600–6. doi:10.7326/0003-4819-153-9-201011020-00010.

5. Observational medical outcomes partnership common data model specifications version 4.0. In: Implementation common data model CDM specifications V4.0. Reagan-Udall Foundation. 2012. http://omop.org/CDM. Accessed November 28 2013.
6. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60. doi:10.1136/amiajnl-2011-000376.
7. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP common data model for active drug safety surveillance. Drug Saf. 2013;36(2):119–34. doi:10.1007/s40264-012-0009-3.
8. Schlienger RG, Jick H, Meier CR. Use of nonsteroidal anti-inflammatory drugs and the risk of first-time acute myocardial infarction. Br J Clin Pharmacol. 2002;54(3):327–32.
9. Garcia Rodriguez LA, Perez Gutthann S. Use of the UK general practice research database for pharmacoepidemiology. Br J Clin Pharmacol. 1998;45(5):419–25.
10. Matcho A. OMOP CDM v4.0 CPRD ETL version 8. In: Implementation common data model CPRD by Janssen research and development. Reagan-Udall Foundation. 2013. http://omop.org/CDM. Accessed Nov 28 2013.
11. Reich CRP, Torok D, Vereshagin S, Khayter M, Welebob E. Standard vocabularies in observational data analysis version 4.0. In: Implementation vocabularies standard vocabulary specification. 2012. http://omop.org/CDM. Accessed Nov 28 2012.
12. Ryan PB. Establishing a drug era persistence window for active surveillance. In: White papers. 2010. http://omop.org/OMOPWhitePapers. Accessed Apr 4 2014.
13. Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case–control method: lessons for developing a risk identification and analysis system. Drug Saf. 2013;36(Suppl 1):S73–82. doi:10.1007/s40264-013-0105-z.